DOCUMENT RESUME

ED 175 943 TM 009 682

AUTHOR Ivens, Stephen H.
TITLE The Promise of CRT: Panacea or Pandora.
PUB DATE 10 Apr 79
NOTE 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Standards; Achievement Tests; *Criterion Referenced Tests; Historical Reviews; Mastery Tests; Measurement Goals; Norm Referenced Tests; Test Construction; *Testing Problems; Test Interpretation
IDENTIFIERS *Domain Referenced Tests; Educational Panaceas

ABSTRACT

After reviewing the history and evolution of criterion referenced tests and citing the writings of other researchers, the author summarizes the current state of criterion referenced tests. They are used to measure an individual's status with respect to a well-defined behavior domain; the behavior domain can be well-ordered, structured, or sequential; domain referenced tests are also criterion referenced; criterion referenced tests are not necessarily also domain referenced; specified performance standards are not integral to criterion referenced tests; criterion referenced tests will not replace norm referenced tests; and criterion referenced tests are a panacea, rather than a Pandora's box. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Stephen H. Ivens

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED175943

TM009 682

The Promise of CRT: Panacea or Pandora[1]

Stephen H. Ivens
The College Board

Criterion-referenced testing (CRT), originally proposed by Glaser (1963), popularized by Popham and Husek (1969), and a topic of extensive research, discussion, and debate ever since, has come under increased scrutiny by the parties at interest as a result of the current furor over minimum-competency testing and the setting of standards for grade-to-grade promotion and high school graduation. The CRT panacea, suggested by many in the late 60's and early 70's as an alternative to, or possibly the replacement for, norm-referenced testing (NRT), however, apparently has not yet materialized -- at least according to last year's highly touted "Presidential Debate," where W. James Popham, one of criterion-referenced testing's most eloquent and influential advocates was unwilling to cite any test in use as an exemplar of CRT (1978a, p. 9).

Positioning a product against the leader in the field is a common practice in advertising -- for example, 7-Up the Uncola, and the Avis' "We're number 2" campaigns. For ten years, at least, criterion-referenced tests have been positioned against norm-referenced tests -- and conceptually the campaign has been successful. But are criterion-referenced tests alive and well, or have we unknowingly pulled a coup that will go down in the annals of Madison Avenue -- created and sold to the educational community a product that may not exist?

---

1. A paper presented at an AERA/NCME symposium "A Criterion-Referenced Approach to the Measurement of Reading Comprehension," San Francisco, April 10, 1979.

I will answer this question later, but first let me admit up front that I do not intend to offer many new ideas today. Rather, I will recount a bit of history, borrow some words of others, and clarify a few confusions regarding criterion-referenced testing to answer the question in the title of this presentation: Panacea or Pandora.

If one recounts history in a strict chronological fashion there is a risk that all of the original confusions will remain. Therefore, my history will start with the Winter 1978 issue of the Review of Educational Research and the lead article by Hambleton and others (1978). This excellent article, which should be mandatory reading for those interested in criterion-referenced testing, adopts Popham's 1975 definition of criterion-referenced tests: "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavior domain (Hambleton, et al., 1978, p. 2)." The article continues with the observation, however, that this definition is similar to definitions given by many for a "domain-referenced test." The authors then conclude that "the term 'domain-referenced test' is more descriptive of what we are doing. However, we agree with W. James Popham that there is considerable public support currently for criterion-referenced tests and that it would be unfortunate if we began a new campaign for domain-referenced tests (p. 2)."

Well, so much for Pandora -- she is alive and well. We are told that criterion-referenced tests are really domain-referenced tests -- that the public wants the former, but we've got the latter -- and it would

be unfortunate to begin a new campaign. "Unfortunate" -- what a curious word to use in this situation.

Equally curious, if criterion-referenced tests are really domain-referenced tests, is last year's debate between Ebel and Popham. Did Popham, ever quick to dismiss "straw person allegations" in the debate, actually create a straw person opponent in Mr. Ebel? Would Ebel have supported criterion-referenced tests if he knew that they were really domain-referenced tests? The question is rhetorical, but I ask only because Hambleton et al. (1978) stated that "perhaps the best examples of domain specifications to date are those prepared by Ebel (1962) and Hively, Patterson, and Page (1968) (p. 33)." I will return to this current history, but let me digress now to 1962 with this reference to Ebel's work which was entitled Content standard scores.

I am sure it was coincidental, but 1962 was the year that Glaser and Klaus first used the term "criterion-referenced tests." A subsequent article by Glaser in 1963, however, is more important in terms of defining criterion-referenced tests in relation to the measurement of educational achievement. The definition by Glaser, so often taken out of context in the past to support positions that may in fact have been untenable is as follows:

> "Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance.

> An individual's achievement level falls at some point on this continuum as indicated by the behaviors he (sic) displays during testing. The degree to which his (sic) achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum. A student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correpondence between what an individual does and the underlying continuum of achievement (1963, p. 519-520)."

Parenthetically, I cannot resist noting that in this article, Glaser also referred to Ebel's 1962 article which proposed similar ideas -- Ebel that same apparent arch-enemy of criterion-referenced tests who particpated in the debate.

Any recounting of the history of criterion-referenced testing must, unfortunately, treat the issue of standards -- an issue which is currently on the minds of many and often discussed in the context of criterion-referenced testing.

Following Glaser's 1963 article, the world of criterion-referenced testing was relatively quiet until the Popham/Husek article in 1969 hit the streets. This article defined a criterion-referenced test as one "used to identify an individual's status with respect to an established standard of performance (p. 1)." Although the article, read in its entirety, did not advocate a single performance standard, the die was cast. Mastery, criterion performance, and the separation of students into have's and have not's became an inseparable part of criterion-referenced testing for many proponents. Glaser and Nitko in 1971, again taken out of context, stated that "a criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards (p. 653)." "Specified performance standards" from Glaser and Nitko reinforced Popham and Husek's "established standard of performance" and people soon forgot the original definition for criterion-referenced tests -- or at least they forgot that Glaser emphasized that "the standard ... is the behavior which defines each point along the achievement continuum." Note again -- "the standard ... is the behavior which defines each point along the achievement continuum."

Professionally we have come full circle -- the current reference to a "domain of behaviors" is not unrelated to Glaser's 'behavior which defines each point on the achievement continuum." But the early to mid-1970's philosophy that criterion-referenced tests compare students to

an absolute standard, which is good, rather than to each other, which is bad, has left its mark -- or should I say scars.

In emphasizing the fact that criterion-referenced testing and standards are conceptually and pragmatically separate, I would be remiss if I did not acknowledge that criterion-referenced tests, by most any definition, can be used to set and assess standards. But then colleges have used the SAT for setting and assessing standards for years, and no one has yet labeled the SAT a criterion-referenced test.

At this point in an earlier draft of my remarks, a colleague asked "What are you going to do now? You have picked on domain-referenced tests; implied that current definitions of criterion-referenced tests are wrong; confused those who hope that criterion-referenced tests will be the alternative to norm-referenced tests; and, most of all you have surely offended Mr. Popham." Clearly, my colleague was perceptive -- I am probably guilty on all counts, and I still have five minutes to go.

The Popham (1975) definition of criterion-referenced tests, which says that such tests are "... used to ascertain an individual's status with respect to a well-defined behavior domain (p. 130)" presents no prescription on the construction of such tests. Neither did Glaser in 1963. Both place emphasis on the interpretation of the results of such tests rather than on the construction of the instrument per se.

In practice, however, some people have argued that in order for criterion-referenced interpretations to be made, the tests must be constructed in certain ways. This is where the domain-referenced tests

come in, even though the proponents do not completely agree on the test construction process. On the one hand there is the item-form approach of Hively, Patterson, and Page (1968), Bormuth (1970) and Kriewall (1969) to name three which are most prescriptive and restrictive. On the other hand, Popham (1978), Millman (1974) and Hambleton and Eignor (1978) permit considerably more latitude in their construction of domain-referenced tests. Regardless of which procedure one follows, one can produce instruments that yield criterion-referenced interpretations. The theory is sound and well-developed for domain-referenced tests and such tests can certainly be labeled criterion-referenced.

There are other kinds of tests, however, that can also be labeled criterion-referenced -- tests that may not be domain-referenced in today's use of the term. Popham and Husek acknowledged in their 1969 article that two types of criterion-referenced tests were possible: one was what we now refer to as domain-referenced -- the other was the familiar in theory, hard to find in practice, Guttman-type scale. The Guttman-type scale is clearly criterion-referenced by Glaser's definition which used the achievement continuum concept, but may not be a domain-referenced test in the current sense of the term. As Gray noted in the Spring 1978 issue of the Review of Educational Research, the original achievement continuum definition lost out to the domain description. But this does not have to be the case. There is room for both, and possibly others, and there is a criterion-referenced test that is alive and well based

on the achievement continuum concept of score interpretation rather than on the domain description (Koslin, 1979).

In summary, then, where are we -- or more precisely, where am I leaving you?

1. Is a criterion-referenced test used to ascertain an individual's status with respect to a well-defined behavior domain? - Yes.

2. Can the behavior domain be continuous -- that is well ordered, structured, or sequential? - Yes.

3. Are domain-referenced tests also criterion-referenced? - Yes.

4. Must a test be domain-referenced to be criterion-referenced? - No.

5. Are specified performance standards intergral to criterion-referenced tests? - No.

6. Will criterion-referenced tests replace norm-referenced tests? - No.

7. Criterion-referenced tests: Panacea or Pandora? - Panacea.

Oh, what about the fact that I offended Mr. Popham? Never fear -- the seven answers all come from his 1969 article.

REFERENCES

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Ebel, R. L. Content standard scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. Gagne (Ed.), Psychological principles in systems development. New York: Holt, Rinehart, & Winston, 1962.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington: American Council on Education, 1971.

Gray, W. M. A comparison of Piagetian theory and criterion-referenced measurement. Review of Educational Research, 1978, 48, 223-249.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1978.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hively, E., Patterson, H. L., & Page, S. A. A 'universe-defined' system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Koslin, B. L. The domain of reading comprehension and it's measurement. Paper presented at an AERA/NCME symposium, "A Criterion-Referenced Approach to the Measurement of Reading Comprehension." San Francisco, April 1979.

Kriewall, T. E. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan Publishing Co., 1974.

Popham, W. J. Educational evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Popham, W. J. The case for criterion-referenced measurements. Educational Researcher, 1978a, 7 (11), 6-10.

Popham, W. J. A lasso for runaway test items. Paper presented at the first annual Johns Hopkins University national symposium on educational research, "Criterion-Referenced Measurement: The State of the Art," Washington, October 1978b.

Popham, W. J. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.